Pearson Estimators in Regression Analysis

George P. Mitsopoulos
Athens University of Economics and Business
Department of Economics
and
Ministry of Economics and Finance
Financial and Economic Crime Unit Service

Michael A. Magdalinos[1]
Athens University of Economics and Business
Department of Economics

April 2003

A b s t r a c t

This paper derives an adaptive partial solution for the maximum likelihood normal equations in a regression, under the assumption that the errors belong to the Pearson family. This estimator can be "robustified" producing a M-estimator with satisfactory efficiency for a wider range of error distributions. Monte-Carlo evidence indicates that the finite sample efficiency of this estimator is comparable to the efficiency of the non-parametric scoring estimator. From a computational point of view, however, the proposed technique is much more efficient: The estimator works very well with pre-selected trimming parameters, and the required improvement term can be computed using an auxiliary least squares regression.

[1] *To the memory of professor Michael Magdalinos who was co-author of this paper and died on 18-08-2002*

Contents

1. Introduction

It is often asserted that the Least Squares (LS) estimator loses efficiency drastically in the presence of even small departures from the normality assumption of the error term [see, e.g., Hampel et al. (1986), p. 309, and the references cited therein]. Several nonparametric and semi-nonparametric techniques have been used to recover some of the efficiency loss.

The Non-Parametric Scoring (NPS) estimators use the LS residuals to find a kernel estimator of the error density, which is used in the one-step scoring method to increase efficiency [Stone (1975), Bickel (1982), Bickel and Ritov (1987), Kreiss (1987)]. The Semi-Nonparametric (SNP) technique approximates the error distribution by the product of a Hermite (or similar) series and a normal density, and derives an approximation to the Maximum Likelihood (ML) estimator [Galland and Nychka (1987), Galland and Tauchen (1989), Magdalinos (1993)].

Both the NPS and the SNP estimators are adaptive. That is, the slope parameters are estimated with full asymptotic efficiency, provided that the error density approximation improves as the sample size tends to infinity. In finite samples, however, there are many problems related with the choice of the bandwidth, or the choice of the order of the approximating polynomial. Hsieh and Manski (1987) found that the NPS estimator with preselected bandwidth can be very inefficient, and they suggest the use of data-based procedures (like bootstrapping or cross validation) for the selection of the bandwidth. Pagan and Schwert (1990) report several cases were nonparametric and semi-nonparametric methods yield inferior out-of-sample forecasts.

Here we propose an alternative estimation strategy: Use a general system of distributions to approximate the error density, and then robustify the resulting estimator. If the system is sufficiently rich, then it will contain distributions in the neighborhood of the true error distribution and the estimator will be nearly efficient. The obvious candidate for this job is the Pearson System of Distributions (PSD) [see, e.g., Eldetron and Johnson (1969), Kendall and Stuart (1977), Johnson and Kotz (1970)]. The family includes a wide range of distributional shapes, such as (Normal, Student's t, Beta, Gamma, etc.), and it is parsimoniously parametrized in terms of its first four cumulants. This is rather convenient, as in most practical situations it is unlikely to obtain reliable sample information for the higher order cummulants.

The definition of the PSD is given in terms of the derivative of the log density function. This implies that the subset of the score vector corresponding to the regression parameters can be derived without the explicit identification of the error distribution. An alternative strategy would be to identify a specific Pearson distribution and then to use this distribution to derive the likelihood and the score vector [see, e.g., Lye and Martin (1993)]. However, many interesting distributions belong to the subset (of measure zero) separating the main types. Consequently, such an identification based on sample information only, can be very unstable.

The rest of the paper is organized as follows. In Section 2 we assume that the true error distribution belongs to the PSD, and we derive the one-step scoring estimator for the structural parameters. In Section 3 we drop the assumption of Pearson family errors and we robustify the scoring estimator. Section 4 presents an experimental study of the proposed estimators. Final remarks and suggestions for some straightforward generalizations are made in the concluding Section 5.

2. Regression with Pearson Family Disturbances

Consider the linear regression

$$y_t = \beta_0 + x_t'\beta_1 + \sigma u_t = z_t'\beta + \sigma u_t \ , \tag{1}$$

where

$$z_t' = (1,\ x_t'), \quad \beta' = (\beta_0, \beta_1'), \tag{2}$$

and y_t is the t^{th} observation of the dependent variable, $x_t' = (x_{t1},...,x_{t(n-1)})$ is the vector of "explanatory" variables. The $n \times 1$ vector of structural parameters β is unknown and $\sigma > 0$.

We assume that the disturbances u_t, $t = (1,\ldots,T)$ are independently and identically distributed with density $f(u)$ belonging to the PSD. Further, we assume that the variables in x_t are (weakly) exogenous [see Engle et al. (1983)], so that there is no loss of information in conditioning upon the distribution of x_t.

If the errors in (1) have finite mean and variance, then without loss of information we can assume that the disturbances u_t are standardized, having mean zero and variance one. Then, the equation defining the Pearson system can be written as

$$\frac{d\ln f(u)}{du} = \frac{u-\alpha_1}{\alpha_1 u + \alpha_2\left(u^2-3\right)-1} \equiv \eta(u) \tag{3}$$

where

$$\begin{aligned} \alpha_1 &= -\gamma_1(\gamma_2+6)/A, \\ \alpha_2 &= -\left(2\gamma_2 - 3\gamma_1^2\right)/A, \\ A &= 10\gamma_2 - 12\gamma_1^2 + 12, \end{aligned} \tag{4}$$

and γ_1, γ_2 are the third and fourth cummulants of the distribution $f(u)$, known as the skewness and kurtosis coefficients, respectively. Substituting (4) in (3), we find

$$\eta(u) = \frac{\gamma_1(\gamma_2+6) + 2\left(5\gamma_2 - 6\gamma_1^2 + 6\right)u}{3\gamma_1^2 - 4(\gamma_2+3) - \gamma_1(\gamma_2+6)u - \left(2\gamma_2 - 3\gamma_1^2\right)u^2}. \tag{5}$$

The conditional likelihood of the t observation is

$$\ell(y_t|x_t,\beta,\gamma) = \frac{1}{\sigma} f(u) \tag{6}$$

where $\gamma' = (\sigma, \gamma_1, \gamma_2)$ is the vector of the error parameters. Consequently the (conditional) loglikelihood of all observations is

$$L = -T\ln(\sigma) \sum_{t=1}^{T} \ln f(u_t) \ . \tag{7}$$

Formula (7) is not operational, as it depends upon the unknown functional form of the density $f(u)$. However, assuming that γ is known and using (3) we can find the score vector corresponding to the structural parameters.

$$\frac{\partial L}{\partial \beta} = \sum_{t=1}^{T} \frac{d\ln f(u_t)}{du_t} \frac{\partial u_t}{\partial \beta} = -\frac{1}{\sigma} \sum_{t=1}^{T} \eta(u_t) z_t \tag{8}$$

Let $\hat{\beta}$, $\hat{\sigma}$ be the LS estimates, and $\hat{u}_t$ the standardized LS residuals. It is easy to show that the skewness and kurtosis coefficients are consistently estimated by

$$\hat{\gamma}_1 = \frac{\sum_{t=1}^{T} \hat{u}_t^3}{T}, \quad \hat{\gamma}_2 = \frac{\sum_{t=1}^{T} \hat{u}_t^4}{T} - 3 \ . \tag{9}$$

Substituting (9) in (5) we find $\hat{\eta}(u)$. The Newton-Raphson method for the solution of (8) suggests the one-step improvement procedure

$$\tilde{\beta} = \hat{\beta} + k\hat{\sigma}\hat{c} \tag{10}$$

where,

$$\hat{c} = \left[\sum_{t=1}^{T} \hat{\eta}(\hat{u}_t)^2 z_t z_t' \right]^{-1} \sum_{t-1}^{T} \hat{\eta}(u_t) z_t \tag{11}$$

the correction vector and $k \in (0, 1)$ a constant witch will be discussed in the section 4.

We shall refer to $\tilde{\beta}$ as the Pearson Improved (PI) estimator. If the error parameters γ were known, then the estimator $\tilde{\beta}$ would be asymptotically fully efficient [LeCam (1956)]. The use of $\hat{\gamma}$ causes inefficient estimation of the regression constant β_0, which is, partially, an error parameter absorbing the effects of a potentially non-zero error mean.

THEOREM 1. If the error distribution has finite moments up to the fourth order and belongs to the PSD, then the PI estimator $\hat{\beta}_1$ of the slope parameters is asymptotically efficient, with asymptotic covariance matrix

$$V(\tilde{\beta}_1) = (J_{11} - J_{10}J_{01}/J_{00})^{-1} \tag{12}$$

where

$$J_{ij} = -E\left(\partial^2 L / \partial\beta_i \partial\beta_j\right), \quad (i,j = 0,1).$$

Proof : Manski (1984) shows that, under general conditions which include our assumptions as a special case, the Stein (1956) condition is satisfied. Consequently, the solution of

$$\partial L(\beta, \gamma)/\partial \beta = 0, \tag{13}$$

produces asymptotically efficient estimators of the slope parameters. The asymptotic equivalence of the PI estimator with the solution of (13) completes the proof of the theorem.

Notice that the lower $(n-1)\times(n-1)$ submatrix of

$$\hat{V} = \sigma\left[\sum_{t=1}^{T} \hat{\eta}(\hat{u}_t)^2 z_t z_t'\right]^{-1} \tag{14}$$

is a consistent estimator of the matrix (12). Consequently, (14) can be used to construct pivotal test statistics for the slope parameters.

3. Robustification of the improvement procedure

Since in the calculation of the PI estimator has been used $\hat{\eta}(\hat{u}_t)$ to approximate the unknown quantity $\eta(u_t)$, the small sample efficiency of the estimator depends on how well the distribution of $\hat{\eta}(\hat{u}_t)$ imitates the distribution of $\eta(u_t)$. For large samples ($T > 100$, say) the distribution of $\hat{\eta}(\hat{u}_t)$ is sufficiently close to that of $\eta(u_t)$ and the PI estimator is nearly efficient. In small samples, however, there is considerable discrepancy between the two distributions.

We compare the typical shape of the distribution of $\eta(u_t)$ (line L1) with that of $\hat{\eta}(\hat{u}_t)$ (line L2) for $T = 30$ and $T = 50$, (Figures 1, 2). The theoretical and estimated values were produced from a simple model with errors following a t distribution with five degrees of freedom (see Section 4).

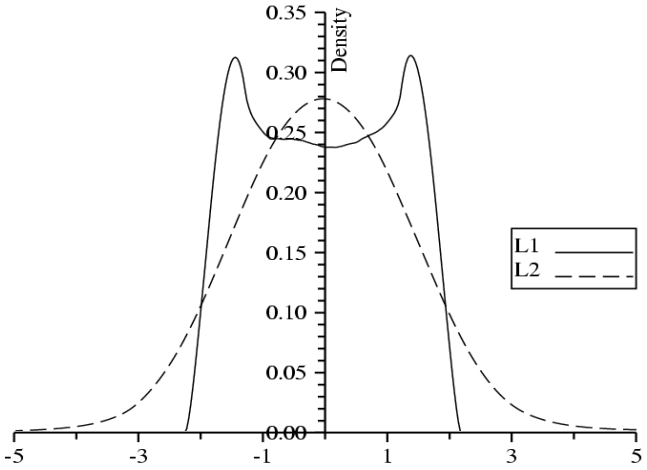

Figure 1. Kernel estimates of the densities of $\eta(u_t)$: L1, $\hat{\eta}(\hat{u}_t)$: L2 for the distribution t with five degrees of freedom, on the sample size $T = 30$.

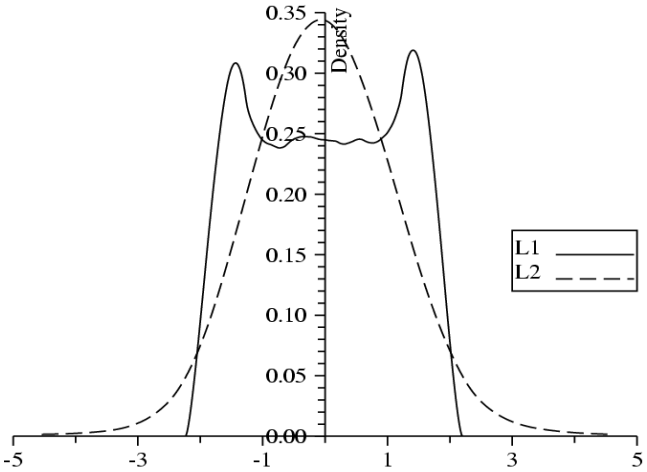

Figure 2. Kernel estimates of the densities of $\eta(u_t)$: L1, $\hat{\eta}(\hat{u}_t)$: L2 for the distribution t with five degrees of freedom, on the sample size $T = 50$.

Figure 1 and 2 shows that the estimation of $\eta(u_t)$ has a "normalizing" effect upon its distribution. In particular, the tails become fatter and longer, and the flat area in the center of the distribution disappears. It can be conjectured that this may be caused by the inefficiency in the estimation of γ_1 and γ_2. We have tried several alternative estimation techniques, such as Huberizing or Wisconsizing the LS residuals, using the Pukelsheim (1980) estimators, etc. Some of them improved the estimation of γ's, but they did not reduce significantly the discrepancy between the distributions of $\eta(u_t)$ and $\hat{\eta}(\hat{u}_t)$. It seems that the problem

originates from the use of the same set of data to estimate both the structural and the error parameters. This is against Bickel's (1982) advice, but it is unavoidable when the sample size is rather small.

A better way to robustify the improvement procedure is by operating directly on the distribution of $\hat{\eta}(\hat{u}_t)$. Given two positive constants $c_2 > c_1 > 0$, we define the function

$$\varphi(u) = \begin{cases} sgn[\eta(u)]c_1 & if\ |\eta(u)| < c_1 \\ \eta(u) & if\ c_1 \le |\eta(u)| \le c_2 \\ sgn[\eta(u)]c_2 & if\ |\eta(u)| > c_2 \end{cases}, \tag{15}$$

where $sgn(\eta)$ is the sign of η. The distribution of $\hat{\varphi}(\hat{u}_t)$ imitates the distribution of $\eta(u_t)$ better than that of $\hat{\eta}(\hat{u}_t)$. The choice of c_1 and c_2 will be discussed in the next section.

Using $\hat{\varphi}(\hat{u}_t)$ instead of $\hat{\eta}(\hat{u}_t)$ in (11) gives the Pearson M-estimator, denoted by PM,

$$\overline{\beta} = \hat{\beta} + \hat{\sigma}\left[\sum_{t=1}^{T} \hat{\varphi}(\hat{u}_t)^2 z_t z_t'\right]^{-1} \sum_{t-1}^{T} \hat{\varphi}(u_t) z_t \tag{16}$$

The estimator (16) can be interpreted as the first step in the Newton-Raphson iterations for the solution of the equations

$$-\frac{1}{\sigma}\sum_{t=1}^{T} \varphi(u_t) z_t = 0\,. \tag{17}$$

Consequently, the estimator (16) is asymptotically equivalent with the solution of (17) which is an M-estimator [Hampel et al. (1986) p. 315].

We define the matrices

$$A = E[zz'],\ M = E[\varphi'(u)zz'],\ Q = E\left[\varphi(u)^2 zz'\right]. \tag{18}$$

Since $\varphi'(u) = \partial\varphi/\partial u$, and $[\varphi(u)]^2$ are bounded almost everywhere, the matrices M, Q are finite (non-singular) if, and only if, the matrix A is finite (non-singular).

THEOREM 2. If A is non-singular, then the influence function of the estimator (15) *is*

$$IF(\bar{\beta}) = \varphi(u)M^{-1}z \ . \tag{19}$$

Proof : Following Marona and Yohai (1981) we can show that (19) is the influence function of the solution of (17). The theorem is implied from the asymptotic equivalence of (16) and (17).

As the influence function (19) is bounded for $u \in P$, the PM estimator is robust with respect to misspecification of the error distribution. Thus, it is expected to work satisfactorily even for error distributions not belonging to the PSD.

Moreover, the results of Yohai and Marona (1979) imply that the asymptotic variance matrix of $\bar{\beta}$ is

$$V = \sigma^2 M^{-1} Q M^{-1} . \tag{20}$$

Pivotal quantities calculated from consistent estimators of V lead to asymptotically valid tests. In particular, we may use the estimates

$$\hat{M} = \sum_{t=1}^{T} \hat{\varphi}(\hat{u}_t) z_t z_t' / T \ , \tag{21}$$

$$\hat{Q} = \sum_{t=1}^{T} \hat{\varphi}(\hat{u}_t)^2 z_t z_t' / T \ . \tag{22}$$

4. Sampling experiments

To examine the relative efficiency of the PI and the PM estimators in samples of small or medium size Monte-Carlo experiments were carried out. The data generating model is a simple version of equation (1), that is

$$y_t = 15.0 + 7.0x_t + 15.0u_t, \ (t = 1,...,T) \tag{23}$$

where in every experiment x_t was generated as independent, uniform $(-1,\ 1)$ variable, and the disturbance term u_t was generated from one of distributions given in Tables 1 and 2.

Table 1.
Distributions that belong to the PSD, with finite $\gamma's$.

	Distribution		μ	σ	γ_1	γ_2
1.	N(*0, 1*)	Standard Normal	0.000	1.000	0.000	0.000
2.	t(*5*)	Student t with $v = 5$ d.f.	0.000	1.291	0.000	6.000
3.	B(*0.5, 3*)	Beta with parameters $a = 0.5$, $b = 3$.	0.143	0.165	1.575	2.224
4.	B(*4, 2*)	Beta with parameters $a = 4$, $b = 2$.	0.667	0.178	-0.468	-0.375
5.	G(*0.5, 1*)	Gamma with parameters $a = 0.5$, $b = 1$.	0.500	0.707	2.828	12.000
6.	G(*5, 2*)	X^2, Chi square with $v = 10$ d.f.	10.000	4.472	0.894	1.200
7.	F(*2, 9*)	Fisher F with $v_1 = 2$ and $v_2 = 9$ d.f.	1.286	1.725	5.466	146.444

Table 2.
Distributions that belong to the PSD, with no finite $\gamma's$ or do not belong to the PSD

	Distribution		μ	σ	γ_1	γ_2
1.	t(*3*)	Student t with $v = 3$ d.f.	0.000	1.732	-	-
2.	Ca(*0, 1*)	Cauchy with parameters $a = 0$, $b = 1$.	-	-	-	-
3.	LN(*0, 1*)	LogNormal, the distribution of $exp(z)$ where z is distributed as N(*0, 1*).	1.649	2.161	6.185	110.936
4.	W(*2, 1*)	Weibull with parameters $a = 2$, $b = 1$.	0.886	0.463	0.631	0.245
5.	W(*8, 2*)	Weibull with parameters $a = 8$, $b = 2$.	1.027	0.152	-0.534	0.328
6.	VCN	Variance Contaminated Normal $0.9N(0,0.1)+0.1N(0,9)$.	0.000	0.995	0.000	21.821
7.	BSM	Bimodal Symmetric Mixture of two Normals, $0.5N(-3,1)+0.5N(3,1)$.	0.000	3.162	0.000	-1.620

To preserve the comparability of the estimators across different experiments, the pseudo-random numbers were rescaled to have zero mean and unit variance. The empirical moments were used for the Cauchy distribution, and the theoretical moments for all the other distributions.

For each distribution we consider sample sizes of

$$T = (30, 50, 100, 200, 500).$$

which give us a total of 60 experiments. Each experiment consists of $10{,}000$ replications and it was executed by a double precision Fortran program. The pseudo-random number were produced by NAG subroutines and by (tested) subroutines written by the authors.

A preliminary set of experiments has been conducted to examine the performance of the PM estimator (16) for different values of the constants c_1 and c_2. The distribution of PM seems to be almost independent from the choice of c_2. All the values in the range 7 to 14 gave very similar results. Relatively more sensitive is the distribution to the choice of c_1. The optimal choice of c_1 differs according to the error distribution, but all the optimal choices belong to the interval $(2, 4)$. Since the bootstrapping (or cross-validation) techniques used to determine the optimum value of c_1 are quite demanding in computer time, we have decided to fix c_1 at the median of the optimum values, that is to take $c_1 = 3$.

From each experiment we produce Monte Carlo estimates of the Mean Square Error (MSE) for the following estimators

- LS, the Least Squares estimator,
- PI, the Pearson Improved estimator (10), with $k = 1$,
- PIk, the Pearson Improved estimator (10), with $k = (1/exp(\|c\|))^2$ and $\|c\|$ the Euclidian norm of the correction vector (11).
- PM, the Pearson M-estimator (15) with $c_1 = 3$ and $c_2 = 9$,
- PR, the PI estimator, with the theoretical values of γ_1, γ_2 used in the calculation of $\eta(\hat{u})$.
- LL, the Linearized Likelihood estimator with the real values of the error parameters substituted in the one-step scoring method.

The last estimator is the "ideal" which both the PI and PM estimators try to imitate. The MSE's for the estimation of the slope parameter $b_1 = 7.0$ are given in Table 3 for the

distributions that obey the assumptions of Theorem 1, while Table 4 gives the MSE's for those that the first two do not have finite γ's, and the last four do not belong to the PSD.

Table 3

Mean Square Error of the estimators for the Parameter $\beta_1 = 7.0$

Distribution	T	OLS	PI	PIk	PM	PR	LL
N(0,1)	30	21.60	83.53	23.48	22.46	21.60	21.60
	50	13.21	19.76	14.44	13.83	13.21	13.21
	100	5.71	6.17	6.00	5.93	5.71	5.71
	200	3.26	3.40	3.38	3.38	3.26	3.26
	500	1.31	1.34	1.34	1.36	1.31	1.31
t(5)	30	20.96	33.66	21.15	19.66	18.12	18.12
	50	13.02	14.70	12.70	12.06	11.01	11.01
	100	5.67	5.31	5.17	5.13	4.72	4.72
	200	3.18	2.78	2.75	2.89	2.60	2.60
	500	1.33	1.10	1.09	1.19	1.06	1.06
B(0.5,3)	30	21.85	15.72	13.99	11.23	10.00	10.00
	50	13.38	9.86	7.64	5.46	6.22	6.22
	100	5.65	4.98	2.74	1.44	2.98	2.98
	200	3.22	3.22	1.64	0.50	2.01	2.01
	500	1.31	1.30	0.83	0.21	0.99	0.99
B(4,2)	30	21.40	229.34	19.10	21.64	23.65	23.65
	50	13.27	50.34	11.36	13.37	13.39	13.39
	100	5.71	12.21	4.62	5.73	4.88	4.88
	200	3.30	4.19	2.59	3.24	2.63	2.63
	500	1.33	1.02	0.93	1.27	0.93	0.93
G(0.5,1)	30	21.77	9.98	13.73	9.11	9.43	9.43
	50	12.41	6.84	7.85	4.31	5.65	5.65
	100	5.55	3.73	3.74	1.52	2.90	2.90
	200	3.21	2.57	2.48	0.70	2.02	2.02
	500	1.31	1.20	1.11	0.24	1.02	1.02
G(5,2)	30	21.66	55.88	19.00	20.39	19.12	19.12
	50	13.07	25.48	11.10	12.14	10.96	10.96
	100	5.72	8.33	4.63	5.10	4.15	4.15
	200	3.27	3.13	2.59	2.87	2.25	2.25
	500	1.30	0.99	0.99	1.13	0.85	0.85
F(2,9)	30	22.52	12.97	15.72	10.28	17.22	17.22
	50	13.23	9.26	9.87	5.93	10.08	10.08
	100	5.62	4.87	4.73	2.72	4.27	4.27
	200	3.21	3.04	2.96	1.77	2.43	2.43
	500	1.35	1.36	1.33	0.89	1.03	1.03

Table 4

Mean Square Error of the estimators for the Parameter $\beta_1 = 7.0$

Distribution	T	OLS	PI	PIk	PM	PR	LL
Ca(0,1)	30	22,76	26,58	17,03	11,95	-	8,25
	50	13,47	17,37	9,74	5,97	-	4,22
	100	5,71	10,21	4,29	2,14	-	1,55
	200	3,28	10,26	2,76	1,23	-	0,83
	500	1,30	6,20	1,30	0,67	-	0,32
t(3)	30	20.88	26.04	19.04	16.70	-	14.51
	50	12.97	13.45	11.32	10.12	-	9.00
	100	5.81	5.26	4.75	4.24	-	6.11
	200	3.26	2.74	2.51	2.50	-	3.04
	500	1.33	1.07	0.97	0.99	-	1.55
LN(0,1)	30	23.00	12.44	15.65	9.33	14.11	12.70
	50	12.69	9.08	9.75	5.52	9.10	8.76
	100	5.66	4.93	4.94	2.63	4.45	5.12
	200	3.23	3.06	3.05	1.76	2.85	3.65
	500	1.34	1.32	1.32	0.86	1.28	2.40
W(2,1)	30	21.93	70.34	19.48	21.53	19.29	21.36
	50	13.75	29.29	11.61	13.42	11.06	12.41
	100	5.76	7.62	4.58	5.49	4.07	4.69
	200	3.25	2.81	2.40	3.07	2.10	2.48
	500	1.30	0.83	0.84	1.19	0.76	0.92
W(8,2)	30	22.18	72.26	21.97	22.27	20.61	20.50
	50	13.14	20.91	12.48	12.92	11.58	11.52
	100	5.71	6.37	5.36	5.63	4.98	4.94
	200	3.23	3.06	2.96	3.17	2.82	2.79
	500	1.35	1.21	1.21	1.33	1.18	1.16
VCN	30	21.68	31.85	16.53	12.11	20.74	10.16
	50	13.08	21.92	9.87	7.21	16.84	4.73
	100	5.60	11.82	3.84	2.76	10.49	1.46
	200	3.18	7.30	1.86	1.32	7.31	0.64
	500	1.32	3.15	0.64	0.43	3.63	0.21
BCN	30	21.79	108.74	19.30	16.82	16.51	11.26
	50	13.40	27.33	12.40	8.75	11.49	5.45
	100	5.47	6.13	5.20	2.94	5.14	1.45
	200	3.26	3.17	3.18	1.69	3.17	0.60
	500	1.33	1.31	1.32	0.70	1.31	0.17

From Tables 3 and 4 we observe that the MSE of the LS estimator is (approximately) the same for all the distributions of the error term. Non-normality implies an efficiency loss relatively to the LL estimator, but not an absolute efficiency loss. It seems that the LS estimator fails to utilize the additional information in the non-normal errors [see Rao (1973) p. 162 for the proof of the minimum information content of the normal distribution]. In absolute terms, however, the LS estimator is equally efficient for normal and non-normal errors.

The distributions of PR and LL estimators are identical for PSD errors and almost identical for unimodal non PSD errors (LN, W, VCN). Significant differences exist only for the bimodal BCN, which is not surprising as all the PSD distributions are unimodal. The PR estimator offers a very simple alternative for fitting regressions when the errors are unimodal with known skewness and kurtosis.

The efficiency of the PI estimator is satisfactory for sufficiently large sample size. In most distributions it is more efficient than LS when $T > 100$. Unfortunately, there are cases [N(*0, 1*), B(*4, 2*), W(*2, 1)*, W(*8, 2)*, BCN] where the PI estimator is dramatically inefficient in small samples $(T \le 50)$. We have simply observed that in these cases the discrepancy between the distributions of $\eta(u)$ and $\hat{\eta}(\hat{u})$ is most pronounced.

The PIk estimator improves the efficiency of the PI estimator. This estimator performs almost uniformly better than PI and LS estimators. Moreover the PIk estimator efficiency gains varies up to 52% with respect to the LS estimator, except the case of the normal distribution which is the ideal case for the LS estimator.

The performance of the PM estimator is much better. It dominates the LS estimator almost uniformly and the efficiency gains varies up to 84%. Even in the cases where the LS estimator seems to be better [N(*0, 1*), B(*4, 2*)] the differences are rather small, and most of them are inside the Monte-Carlo error limits. Sometimes the improvement is quite dramatic [B(*0.5, 3*), G(*0.5, 1*), F(*2, 9*), C(*0, 1*), LN(*0, 1*), VCN, BCN]. Very surprising is the fact that in several cases [G(*0.5, 1*), F(*2, 9*)] the PM estimator is more efficient than the LL estimator for all the sample sizes considered. This demonstrates the limitations of the (first order) asymptotic theory when it is used to predict the relative estimator performance in finite samples.

The PM estimator dominates the PIk estimator in the case of the distributions [N(*0, 1*), t(5)] for all $T < 200$ and in the in the case of the distributions [B(*0.5, 3*), G(*0.5, 1*), F(*2, 9*), C(*0, 1*), t(*3*), LN(*0, 1*), VCN, BCN] for all sample sizes. In all the other cases the PIk

estimator is slightly more efficient that PM. Moreover in the cases of the distributions [B(*0.5, 3*), G(*0.5, 1*), F(*2, 9*), C(*0, 1*), LN(*0, 1*), BCN] the improvement is substantial and varies up to 78%. The superiority of PM is more pronounced when we use data-based procedures to determine the optimal value of the constant c_1.

The other distributional aspects of the four estimators (LS, PI, PIk, PM, LL) have also been inspected. All of them seem to be nearly unbiased (typically the estimated bias contributes less than 4% to the estimated MSE), and there is not a systematic ranking of the estimators with respect to the bias. The estimated quantiles and decentiles of their distributions indicate that they are nearly symmetric. Also the distributions of PI, PIk, PM, and LL are rather leptokurtic. Instead of reproducing here the (rather extensive) tables, we prefer to give a sample of the graphs of the estimated densities of the estimators.

Each density was estimated from 10,000 values of the corresponding estimator. We used an Epanechnikov kernel with optimum bandwidth obtained by least-squares cross-validation [Silverman (1986) p. 40-51]. All the estimators are calculated from a simulated sample of size $T = 50$. The first two figures are representative of cases favourable to the PI and PM estimators: The error distributions [B(*0.5, 3*), F(*2, 9*)] have finite fourth order moment and large skewness and kurtosis. It is not surprising that the PM estimator is more concentrated than the LL estimator. The last two figures represent very unfavourable cases: The Cauchy distribution does not have finite moments and the BCN is bimodal, i.e. with shape very different from the shapes in the PSD. Although the LL estimator is more concentrated than the PM estimator, the latter performs very well compared with the LS and PIk estimators.

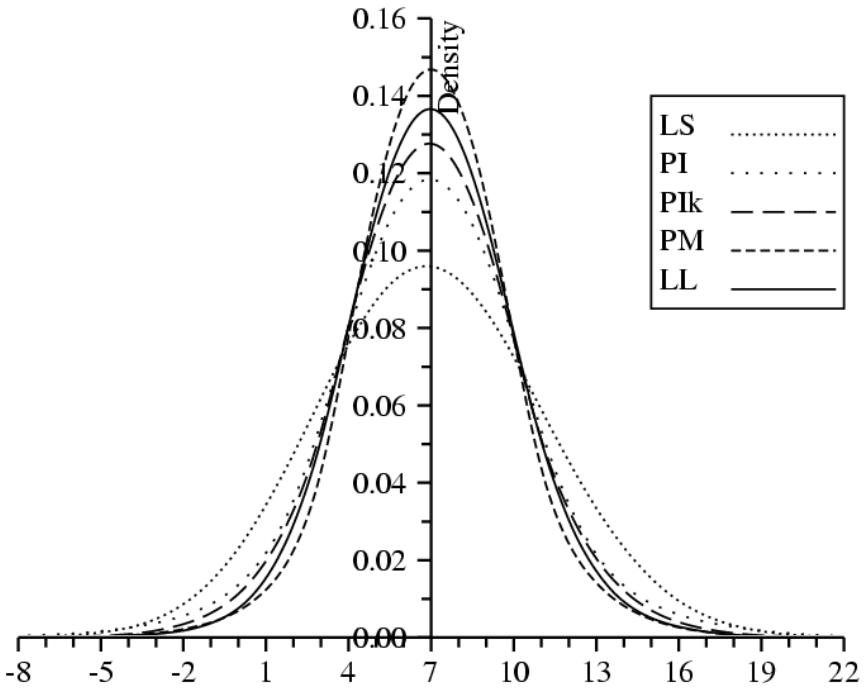

Figure 3. Kernel estimates of the estimator densities for B(*0.5, 3*) disturbances.

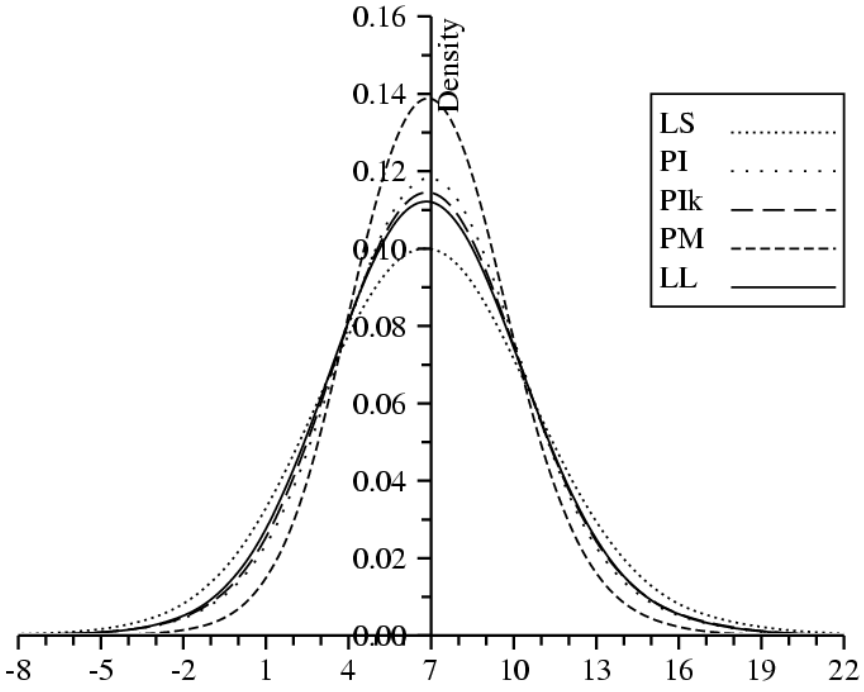

Figure 4. Kernel estimates of the estimator densities for F(*2, 9*) disturbances.

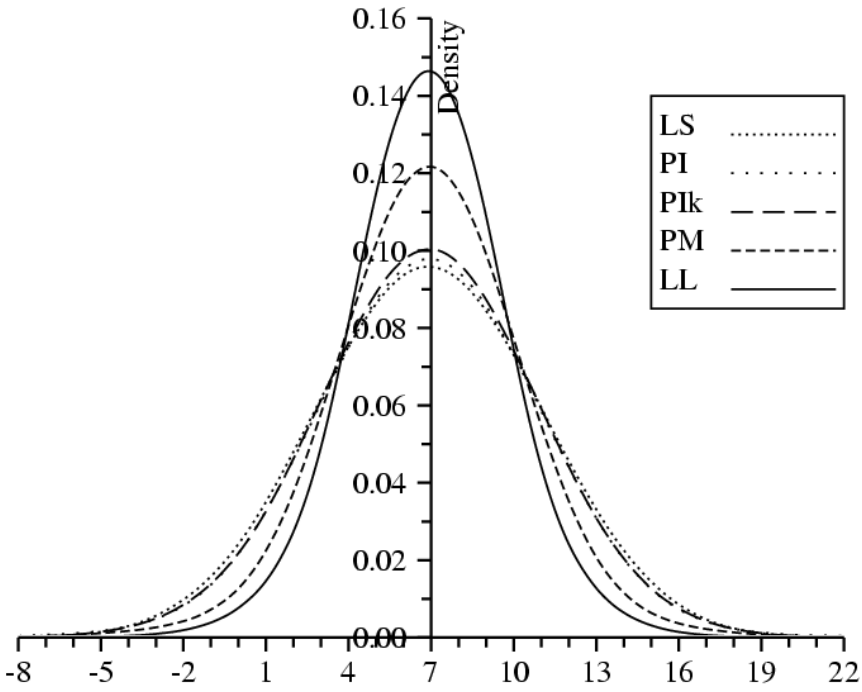

Figure 5. Kernel estimates of the estimator densities for BCN disturbances.

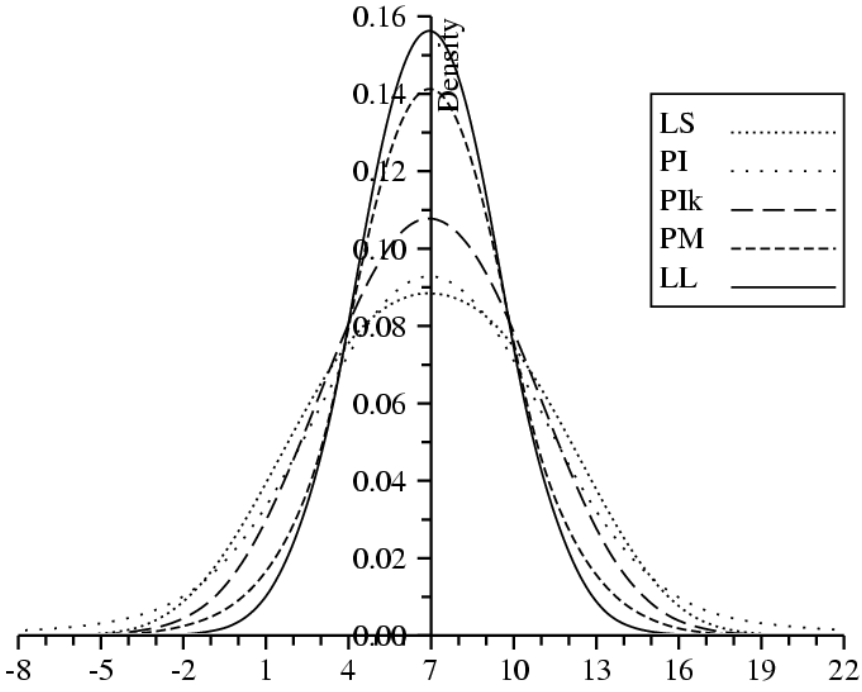

Figure 6. Kernel estimates of the estimator densities for Cauchy(*0, 1*) disturbances.

A second set of experiments was carried out to examine the relative efficiency of the PM and NPS estimators. For this, we reproduced the Hsieh and Manski (1987) experiment for the sample size 50. The model and the data generation process are exactly the same. In Table 5 we report the Root Mean Square Error of the following estimators

- LS, the Least Squares Estimator
- PM, the Pearson M-estimator with constant trimming parameters $c_1 = 3$, $c_2 = 9$.
- BPM, Bootstrap PM with the optimal values of c_1 and c_2 determined by a bootstrap search.
- AML, Approximate Maximum Likelihood with constant trimming and smoothing parameters $t = 3$ and $s = 0.75$ (the median of the optimal values).
- BAML, Bootstrap AML with the optimal values of t and s determined by a bootstrap search.

The error distributions examined are the same as in Hsieh and Manski (1987) (Table 5) and the bootstrap search for each distribution is exactly the same.

Table 5.
Mean Square Error for the Parameter $\beta_1 = 1.0$

Distribution	OLS	PM	BPM	AML	BAML
N(*0, 1*)	0.28	0.29	0.28	0.30	0.31
t(*3*)	0.28	0.27	0.24	0.28	0.24
B(*2, 2*)	0.28	0.29	0.28	0.28	0.29
LN(*0, 1*)	0.28	0.23	0.20	0.50	0.18
VCN	0.29	0.23	0.20	0.72	0.17
BSM	0.29	0.22	0.21	0.60	0.17

Table 5 shows that the PM procedure is more robust than the AML procedure relatively to the choice of trimming and/or smoothing parameters: The maximum improvement of the RMSE's by the bootstrap search is 76% for the AML and 13% for the PM estimator. This is an advantage of the PM estimator which performs satisfactorily with preselected trimming parameters over the AML estimator which requires data based search method for the selection of the bandwidth. Comparison of the RMSE's of the BPM and BAML estimators shows that

their differences are rather small. As expected, the BPM is better for the PSD distributions, whereas the BAML is better for the non-PSD distributions.

5. Concluding remarks

In this paper we propose the use of a M-estimator, which uses the skewness and kurtosis properties of the error distribution to improve the efficiency of the LS estimator. Although the estimator is derived under the assumption of Pearson error distributions, it is robust with respect to this assumption and with respect to the assumption of finite fourth order moments.

The computational requirements are very modest. For example, the improvement (16) can be calculated by regressing a column of 1's on the auxiliary variables

$$\omega_t = \left[\hat{\varphi}\left(\hat{u}_t\right)/\hat{\sigma}\right] z_t \ , \tag{24}$$

using standard statistical software. A preliminary skewness-kurtosis test [Kendall and Stuart (1977)] can be used to detect the nearly normal cases for which the LS estimator is likely to be more efficient than the PM estimator. Data based methods for the selection of the trimming parameter c_1 can be used to increase the efficiency of PM. However, even the rule-of-thumb choice of $c_1 = 3$, $c_2 = 9$ works quite satisfactorily for a wide range of distributions and sample sizes.

Several generalizations of the PM improvement are possible. For example, if the regression is non-linear

$$y_t = \alpha + h\left(x_t, \beta\right) + \sigma u_t \ , \tag{25}$$

where h is a function such that the Non-linear Least Squares (NLS) estimator in (25) is consistent, then the improvement is identical to the improvement in the linear case, provided that we substitute the vector of the regressors z_t' by the vector of partial derivatives

$$d_t' = \left[1, \frac{\partial h}{\partial \beta'}\left(x_t, \hat{\beta}\right) \right], \tag{26}$$

where $\hat{\beta}$ is the NLS estimator of β. Some preliminary experimental results indicate that the efficiency gains over the NLS are of the same order of magnitude as the gains in the linear case.

Finally, the same improvement can be applied to heterogeneously distributed observations. In this case the parameter σ in (1) depends on t, that is $\sigma = \sigma_t, (t = 1,\ldots,T)$ If $\hat{\sigma}_t$ is a consistent estimator of σ_t then the standard GLS transformation results in the equation

$$y_t/\hat{\sigma}_t = (z_t'/\hat{\sigma}_t)\beta + \varepsilon_t \ . \qquad (27)$$

If we assume that the skewness and kurtosis of ε_t are constant, then the improvement process of Section 3 can be applied without change. An alternative approach is to find explicit expressions for the third and fourth moments of ε_t and base the improvement on the derived coefficients of skewness and kurtosis. As the last approach gives rise to several theoretical and practical problems, it will be considered in detail in a future work.

References

Bickel, P.J. (1982). On adaptive estimation. *Annals of Statistics* 10, 647-671.

Bickel, P.J., Ritov, Y. (1987). Efficient estimation in the error variables model. *Annals of Statistics* 15, 513-540.

Eldetron, W.P., Johnson, N.L. (1969). *Systems of frequency curves.* Cambridge University Press.

Engle, R.F., Hendry, D.F., Richard, J.F. (1983). Exogeneity. *Econometrica* 51, 277-304.

Gallant, A.R., Nychka, D.W. (1987). Semi-nonparametric maximum likelihood estimation. *Econometrica* 55, 363-390.

Gallant, A.R. Tauchen, G. (1989). Semi-nonparametric estimation of conditionaly constrained heterogenous processes. *Econometrica* 57, 1091-1120.

Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J. Stahel, W.A. (1986). *Robust Statistics.* John Willey and Sons, New York.

Hsieh, D.A., Manski, C.F. (1987). Monte-Carlo evidence on adaptive maximum likelihood estimation of a regression. *Annals of Statistics* 15, 541-551.

Johnson, N.L., Kotz, S. (1970). *Distributions in statistics: Continuous univariate distributions -1.* John Wiley and Sons, New York.

Kendall, M., Stuart, A. (1977). *The advanced theory of statistics.* Vol. 1, C. Griffin and Co., London.

Kreiss, J.P. (1987). On adaptive estimation in stationary ARMA processes. *Annals of Statistics* 15, 112-133.

Le Cam, L. (1956). On the asymptotic theory of estimation and testing hypotheses. *Proceedings of the 3rd Berkeley Symposium in Mathematical Statistics and Probability 1,* 129-156.

Lye, J.N., Martin, V.L. (1993). Robust estimation nonnormalities and generalized exponential distributions. *Journal of American Statistical Association* 88, 145-194.

Magdalinos, M.A. (1993). Approximate maximum likelihood estimation in linear regression. *Annals of the Institute of Statistical Mathematics* 45, 1, 89-104.

Manski, C.F. (1984). Adaptive estimation of non-linear regression models. *Econometric Reviews* 3, 145-194.

Marona, R.A., Yohai, Y.J. (1981). Asymptotic behaior of general M-estimates for regression and scale with random carriers. *Z. Wahrsch. verw. Geb.* 58, 7-20.

Pagan, A.R., Schwert, W. (1990). Alternative models for conditional stock volatility. *Journal of Econometrics* 45, 267-290.

Silverman, B.W. (1986). *Density estimation for statistics and data analysis. Chapman and Hall,* London.

Stein, C. (1956). Efficient non-parametric testing and estimation. *Proceeding of the 3rd Berkeley symposium in mathematical statistics and probability* 1, 129-156.

Stone, C.J. (1975). Adaptive maximum likelihood estimation of a location parameter. *Annals of Statistics* 10, 647-671.

Pukelsheim, F. (1980). Multilinear estimation of skewness and kurtosis in linear models. *Metrica* 27, 103-113.

Rao, C.R. (1973). *Linear statstical inference and its applications.* 2nd ed. New York: John Wiley and Sons.

Yohai, V.J., Marona, R.A. (1979). Asymptotic behavior of M-estimators for the linear model. *Annals of Statistics* 7, 2, 258-268.